James E. Cooke

The Living Mirror Theory of Consciousness

Abstract: *An explanatory gap exists between physics and experience, raising the hard problem of consciousness: why are certain physical systems associated with an experience of an external world from an internal perspective? The living mirror theory holds that consciousness can be understood as arising from the computational interaction between a living system and its environment that is required for the organism's existence and survival. Maintaining a boundary that protects the system against destructive forces requires an interaction between the organism and its outside world that can be cast in terms of Bayesian inference. The living mirror theory holds that this computational interaction results in statistical properties of the material world that are, in the absence of life, only implicit, becoming explicit in informational terms. This is held to give rise to the beliefs in qualities that constitute consciousness. Consciousness is therefore a necessary feature of all living systems as, in a world governed by the second law of thermodynamics, survival depends on the construction of beliefs regarding the potentially destructive forces in the outside world. From this perspective, consciousness is shown to be not a property of the brain in particular but instead to be a necessary feature of the life process itself.*

1. Why Are Some Physical Systems Conscious?

Why should some physical systems have an internal phenomenal experience of an external world? This is the hard problem of consciousness, a phrase coined by David Chalmers (1995; Chalmers *et*

Correspondence:
Dr. James Cooke, Institute of Behavioural Neuroscience, Department of Experimental Psychology, University College London (UCL), 26 Bedford Way, London, WC1H 0AP, UK. *Email: james.cooke@ucl.ac.uk*

Journal of Consciousness Studies, **27**, No. 9–10, 2020, pp. 127–47

al., 2003). In the paper that introduced the concept of the hard problem, Chalmers writes, 'It is widely agreed that experience arises from a physical basis, but we have no good explanation of why and how it so arises. Why should physical processing give rise to a rich inner life at all? It seems objectively unreasonable that it should, and yet it does' (Chalmers, 1995). With our current theories of physics and consciousness, it is indeed 'unreasonable' to expect that a system such as a human would have an internal experiential character — no account currently exists that satisfyingly draws a link between these two domains. Given that the system is governed entirely by the laws of physics, why couldn't the system operate by blind automation (Nagel, 1970; 1974; Kirk and Squires, 1974; Chalmers, 1996)? To use Thomas Nagel's definition of consciousness (1974), why should it be 'like something' to be a human but presumably not 'like something' to be a rock? Put another way, an explanatory gap currently exists between physics and consciousness (Levine, 1983). If the physical world can be entirely characterized by the laws of physics, why should certain systems be associated with an internal experience of an external world? Why doesn't the operation of systems like ourselves unfold in a blind, mechanistic manner? To date, no solution has been proposed that successfully bridges this gap between mind and matter.

Here I propose the living mirror theory as a solution to the hard problem of consciousness. The living mirror theory holds that phenomenal consciousness arises in the natural world as a result of the entropy-resisting dynamics of living systems. These dynamics are held to result in the emergence of a framework of beliefs in qualities that is argued to be an equivalent description of consciousness. By accounting for the presence of phenomenal consciousness in terms of the thermodyanmic operation of living systems, this theory accounts for how consciousness relates to the physical world and, in doing so, bridges the gap between mind and matter. It can be predicted from this theory that consciousness is not dependent on brains in particular and that all living systems are conscious. In multicellular organisms, the principles relevant for consciousness apply at multiple levels, from individual cells up to brain-wide neural systems. The living mirror theory therefore holds that multiple conscious systems exist within multicellular life forms such as ourselves.

2. The Living Mirror Theory of Consciousness

In contrast to the quantitative material world, consciousness is qualitative. Conscious systems experience qualities ranging from colour to pain; the fact of qualitative experience is the fact of consciousness. Qualities such as colour and pain do not exist in the material description of the world in the absence of conscious systems, yet conscious systems experience these qualities as existing in a wider world. When a red apple is consciously perceived, the conscious experience of colour does not consist only of the quality of redness; it consists of the belief that this quality exists in an outside world. In spite of the absence of qualities such as colour in the material description of the world, certain physical systems manage to construct beliefs in such qualities and, in doing so, become conscious. The hard problem of consciousness can therefore be reformulated as the question of how physical systems come to form beliefs in qualities.

The living mirror theory holds that the computational construction of beliefs in qualities is a necessary feature of the anti-entropic dynamics that define all living systems. Living systems resist the tendency towards disorder by maintaining a boundary with their external environment, and the statistical dynamics of this process have been found to be equivalent to the internal states of the living system instantiating the survival-relevant properties of the external environment (Friston, 2013). The living mirror theory holds that this process results in the implicit statistical structure of the material world becoming explicit in informational terms inside the living system, bringing into existence the beliefs in qualities that constitute consciousness. In order to survive in a universe moving towards increased disorder, all living systems must compute beliefs regarding the qualitative character of the potentially destructive forces surrounding them.

By accounting for consciousness as a necessary computational feature of the thermodynamic operation of living systems, the living mirror theory accounts for how consciousness relates to the physical world, thereby bridging the explanatory gap between matter and consciousness. We live in a universe where certain physical objects create reflections that are made of photons yet correspond to material objects. Similarly, another property of our universe is that living systems necessarily produce informational reflections of the properties of material in their environment in order to survive, accounting for the existence of consciousness.

3. The Physics of Life

The entirety of physical reality can be conceptualized as a single system that tends towards disorder (Figure 1A). The second law of thermodynamics states that the total entropy, or disorder, of an isolated system will increase over time. The totality of physical reality is such an isolated system, and this law describes how the physical constituents of that system inevitably blend into each other as the system moves towards thermodynamic equilibrium. From the perspective of statistical physics, a rock in the ocean is not a system that can be defined as separate from the ocean; it is part of a single larger system whose entropy will increase over time. This results in disarray and intermingling between the rock and the water as the rock gradually erodes and mixes with the ocean. Boundaries therefore do not exist in an isolated system, governed by the second law of thermodynamics.

The defining feature of living systems is that they resist entropy (Schrödinger, 1944). Unlike a rock in the ocean, a single-celled organism consists of parts of physical reality that manage to maintain their combined form. The system operates in such a way as to avoid decay over time, despite the forces external to the living system that produce disorder in non-living systems. The free energy principle characterizes how physical systems produce this behaviour (Friston, 2013). The principle shows that entropy resistance requires parts of the system to form a boundary with parts of the larger system which becomes defined as external by this process. This boundary can be defined as a Markov blanket, a concept from statistical physics that represents the nodes of a networked system that constitute its border (Pearl, 2014). The Markov blanket can be used to identify a system in opposition to its environment (Figure 1B). Crucially, this cannot be done for a pseudo-system such as a rock. The rock exhibits no dynamics that makes it separate from its environment in any meaningful sense.

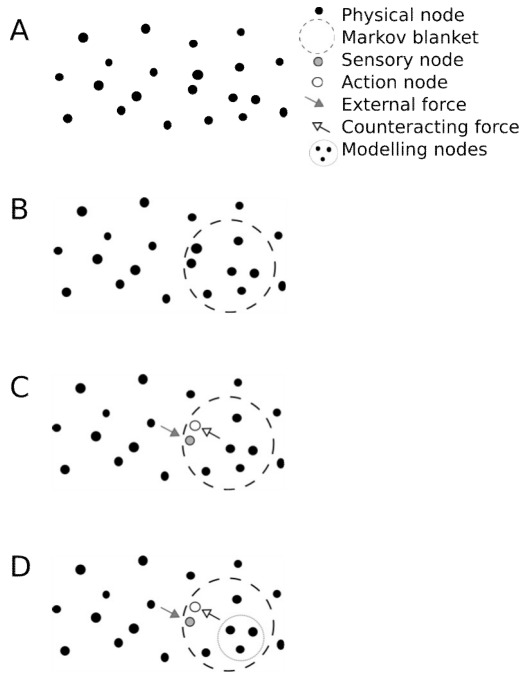

Figure 1. Entropy resistance entails Bayesian inference. A. The totality of physical reality is a single isolated system consisting of physical matter, represented as nodes in a network here. B. Living systems define themselves against the rest of reality by establishing a border in the form of a Markov blanket, a set of nodes that isolate the internal system from the external environment. C. The Markov blanket has sensory nodes that measure the potentially destructive forces that act on the system and active nodes that allow it to undertake behaviours that counteract these forces. In this way the Markov blanket is maintained such that the system within continues to exist. D. The physical dynamics that allow this process to occur can be described as reducing free energy within the system. Free energy reduction is equivalent to parts of the system internal to the Markov blanket acting as a model of the external events that account for the sensory evidence arriving at the Markov blanket. From the perspective of statistical physics, in order for living systems to exist they must infer the properties that exist in their external environment. This is the only way to estimate the nature of potentially destructive forces that might act on the system and thereby resist the fundamental tendency towards disorder (based on Friston, 2013).

4. Form Maintenance Entails Bayesian Inference

In order for a system to exist it must maintain its form. This involves resisting the tendency towards increased disorder over time. In terms of statistical physics, entropy resistance is achieved by maintaining the Markov blanket that bounds the system (Friston, 2013). Markov blanket maintenance is achieved by the blanket having nodes dedicated to measuring the external forces that impinge upon the system and having other nodes dedicated to exerting counteracting forces (Figure 1C). This successfully isolates the internal states of the system from the external environment and produces the capacities for sensation and action in the system respectively. The dynamics of entropy resistance show how sensation and action are necessary features of any living system that exists over time.

The computational action of the system that results in boundary maintenance via sensation and action has been shown to be the result of the system operating in a manner that reduces a statistical quantity called free energy (Friston, Kilner and Harrison, 2006; Friston, 2013). Free energy reduction is equivalent to preventing the entropy of the sensory states of the system from increasing indefinitely, which would lead to the destruction of the system. Systems that exist over time therefore must reduce free energy. The reduction in free energy is also equivalent to performing Bayesian inference (for a quantitative description of the free energy principle, its relationship to Bayesian inference, and a heuristic proof of these principles, see Friston, Kilner and Harrison, 2006, and Friston, 2013). From a Bayesian perspective, the internal states of the system physically instantiate the probabilities of different external environmental causes that could account for the pattern of sensory stimulation that appears on the Markov blanket (Friston, 2013). Successfully predicting future sensory events using this information regarding external events that cannot be known directly allows the system to maintain its form over time by acting in a manner that is in line with its continued existence. In order to take actions in line with its survival, living systems must necessarily infer the properties of the world that account for the sense-data received by the system (Figure 1D).

Recent attempts in neuroscience have been made to explain two aspects of the contents of consciousness, affect and the self model, in terms of free energy minimization. In one recent theory, consciousness is asserted to be the internal experience of the brainstem engaging in homeostasis, a process that can be described in the free

energy framework (Solms and Friston, 2018; Solms, 2018). In this account, consciousness is taken to be the 'surprise' registered in the centrencephalic structures when homeostatic signals are not successfully predicted. In another recent theory, self-consciousness is argued to occur through the inference operations described here operating over a long enough timescale that the system can model the consequences of its own actions and thereby represent itself (Friston, 2018). Affect and self models are appearances within consciousness; however, they are not the fact of consciousness itself.

5. Accounting for the Character of Consciousness

Consciousness is a qualitative space that is private, unified, intentional, and transparent. Intentionality refers to the sense in which consciousness is about the wider world. Transparency refers to the way in which consciousness itself is typically not detected during experience. Instead there is the impression of seeing through consciousness to a world beyond. The living mirror theory holds that entropy-resisting dynamics in bounded systems inevitably produce, at the informational level, an internal framework of beliefs in survival-relevant qualities in the outside world and that this framework of beliefs is an equivalent description of consciousness itself. This equivalence can be seen in the ways in which the living mirror theory accounts for the qualitative, private, unified, intentional, and transparent characteristics of consciousness.

How does the living mirror theory account for the qualitative character of consciousness? The consciously beheld qualitative character of the material world emerges from the interaction of living systems with a physical environment that has no objective qualities but does have implicit statistical structure. The material world can be described at many scales. At the most fundamental level, the entire physical world can be described as a single quantum state; this is all that can be said to exist at this level of description. At the level of atoms, certain particles and forces can be said to exist, and at the level of molecules, certain patterns of interactions between chemicals can be said to exist. Above all of these levels, at the scale at which living systems operate, a rich statistical structure can also be said to exist. The mean kinetic energy in an arbitrarily defined area is a statistical feature of the world, as is the reflectance profile of the same area. An area of molten rock on the ocean floor will have a high mean kinetic energy relative to the surrounding water and it will also preferentially

reflect long wavelengths of light. The kinetic energy and reflectance properties vary together, and this correlation is an example of another statistical feature of the world. Does this statistical structure truly exist, however, in the way that atoms and molecules exist? Once an area has been defined, the mean kinetic energy can be calculated as can any correlations with reflectance properties. The observed statistical structure can be found to be highly reliable, indicating that it truly is a feature of our reality, but it cannot be said to autonomously exist in the absence of the area being defined and the measurement being taken.

Living systems exist in a world rich in statistical structure, some of which will be relevant for the system's survival. This can be seen in the example of the area of molten rock on the ocean floor. The high kinetic energy of the rock relative to the surrounding water has the effect of increasing the entropy of the matter surrounding it, increasing the temperature of the water and mortally increasing the entropy of any living system that ventures too near. Living systems that survive will successfully avoid such areas and thereby keep their free energy minimized and their boundaries intact. The physical evolution of populations of such systems will necessarily endow the successful living systems with the computational capacity to construct beliefs about the nature of these areas of matter so that they can be avoided. The internal physical states of the living system will come to be informative with regards to the relevant properties of the molten rock, such as the kinetic energy and reflectance properties. The system cannot detect these properties directly but receives sensory evidence that is detected by thermal receptors and photoreceptors. These parts of the system's boundary represent relevant quantities — they signal the level of kinetic energy or the wavelength of light detected by the living system. The representation of quantities in the external world is not sufficient for consciousness, however. If this were the case, thermometers would be conscious of temperature as the level of mercury represents the temperature of the local environment. Consciousness cannot be understood as consisting of representations that signal a feature of a separate outside world but must instead be appreciated as an emergent feature of the living system in conjunction with its environment (Rosch, Varela and Thompson, 1991).

In order to survive, living systems cannot merely represent the information they receive, they must compute the relevant properties of the world beyond their boundaries. They must construct an internal framework of survival-relevant qualities that allows them to behave in

ways that increase their survival. Here, we are no longer dealing with the representation of quantities, as it is not possible for living systems to compute an accurate quantitative representation of the relevant properties of the outside world. Instead, according to the living mirror theory, the computational action of living systems results in the relational structure of the world's implicit statistical properties that are relevant to the survival of the organism becoming explicitly instantiated in informational terms by the internal physical states of the system. From this perspective, living systems cast an informational reflection of the qualitative character of the material world that only exists as implicit statistical quantities in the absence of living systems.

How is it possible for qualities to exist in a quantitative world? It is possible for qualities to exist in physical systems at the level of information. Physical states convey information when they signal that the world is one way and not a number of other possible ways the world could be (Tononi, 2004). From this, it can be seen that, while information can be quantified, information is not primarily quantitative. Informativeness can be understood as being relative to a space of possibilities, not a space of quantities. Information is thereby capable of instantiating non-veridical beliefs in qualities when no qualities exist in the quantitative world. Each belief is informative relative to the other possible beliefs in the system. The existence of the belief in red is dependent on the existence of other related beliefs — the experience of an object having a specific colour is informative in that it indicates that the object is not any other colour, rather than indicating certain quantitative properties of the world.

How does the living mirror theory account for the private character of consciousness? According to the living mirror theory, consciousness comes into existence with the asserting of a boundary with the rest of the physical world and is perpetuated by the continued maintenance of this boundary. An internal vs. external duality is brought into existence, and it is through this process that private spaces are brought into existence. Without invoking the boundary that defines the extent of a living system, the private nature of consciousness is difficult to account for. By holding that consciousness is a feature of the life process, the living mirror theory shows that the bounded nature of living systems and the privacy of consciousness are two aspects of the same process of entropy resistance.

How does the living mirror theory account for the unified character of consciousness? Consciousness is unified in the sense that it is the

appearance of a single world, an appearance in which experiences relate to each other despite originating in different sensory channels. The brain is a highly distributed system with no central integration area, raising the problem of how features become bound together to form unified conscious experience (Revonsuo, 1999). This perspective assumes that parallel systems in the brain are independently capable of giving rise to conscious experiences that must be subsequently combined (Damasio, 1989). According to the living mirror theory, the goal of conscious perception is to form useful beliefs about the world. In our universe, the structure of sensory input is often correlated between sensory modalities, making conscious perception an inherently multisensory task. The conscious belief in the presence of a visible human talking to you is not comprised of a visual experience of a face and an independent auditory experience of a voice, with the two experiences being subsequently bound together. Rather, the computational action of the living system results in the construction of the belief in the visible talking human from the start, in which all available sensory evidence is used to build this inherently multisensory belief. From this perspective, no subsequent binding stage need take place (Hohwy, 2013). A second feature of the living mirror theory that accounts for the unity of consciousness is the fact that a living system is defined by a boundary that separates it from a single wider world. Such a system requires a single framework of beliefs in the qualitative character of the external world in order for it to behave in a manner that is in line with its survival. We should therefore not expect consciousness to be fractured. Finally, each belief has no independent existence separate from the entire framework of beliefs, as qualities are informative with respect to the possible space of beliefs, rather than with respect to the quantitative physical world. Beliefs can therefore only exist in a unified framework in which each belief's existence is dependent on the existence of other beliefs.

How does the living mirror theory account for the intentional character of consciousness? The Bayesian models constructed by living systems are beliefs about the properties of the world beyond the living system, making them fundamentally intentional. They emerge as a result of, and are coupled to, the behavioural survival dynamics of the living system — their functional role is to instantiate the survival-relevant properties of the world in a coordinate system that is relevant for behaviour and they must therefore be about the world beyond the living system. This coupling to the outside world is a crucial feature in understanding living systems in thermodynamic terms; the system can

only emerge by constructing beliefs about the outside world. It is this coupling that differentiates beliefs from representations of qualities that have no intentional character.

How does the living mirror theory account for the transparent character of consciousness? Belief in the qualitative character of the world must enable the living organism to successfully engage with the world, in the service of survival. A belief in the presence of food in front of the organism, if functioning correctly, is transparent — only a belief in the belief would subvert this transparency. That is to say, such beliefs lose their transparency when they themselves become represented and, lacking survival value, such meta-representations appear to not have been selected for in living systems (Metzinger, 2009). The lack of meta-representations is all that is required for such beliefs to be transparent and to thereby function appropriately.

In addition to accounting for how consciousness emerges from the operation of the physical world, the living mirror theory also accounts for the qualitative, private, unified, intentional, and transparent characteristics of consciousness. Beyond this, must a theory of consciousness account for what consciousness itself is made of? All scientific descriptions of reality, in both the qualitative and quantitative domains, are ultimately abstract descriptions of how the parts of the system relate to each other and not what the system is ultimately made of. Knowledge proceeds by understanding systems in relational terms — the question of what matter is made of at the fundamental level is impossible to answer, as is also the case for consciousness. The quest to describe essence ultimately proves futile in the case of both matter and mind and so we must confine ourselves to describing systems of interactions in the manner presented here.

6. Implications for the Nature of Consciousness

Complex living systems such as ourselves consist of self-organizsing cells and, as a result, we are a system that is comprised of many subsystems (Kirchhoff *et al.*, 2018). The brain is a system of the kind described here in its own right (Friston, Kilner and Harrison, 2006), as are the individual cells that comprise the brain. It can therefore be predicted from the living mirror theory that islands of consciousness can exist within larger islands of consciousness. For example, there is something that it is like to be the system in the brain that we identify with, but the living mirror theory requires that the individual neurons that comprise that system also have their own conscious experience,

something it is like to be them as they resist thermodynamic equilibrium through the construction of beliefs about their environment. This is the case because each cell follows the same entropy-resisting dynamics as the whole system. It is a bounded system that minimizes free energy through belief construction in order to not dissolve into its surroundings. Within the neuron, mitochondria would also have some experiential character as they construct beliefs about their external world in order to survive. The logical end point of this reasoning is that each cell in our body has an experience of its environment that is separate from and inaccessible to the conscious experience that we identify with. It is important to note that this claim is as consistent with our experience of the world as the claim that individual cells are not conscious, as consciousness is inherently private.

As you read these words, they appear in a particular field of consciousness. A widespread assumption in psychology and neuroscience is that this consciousness is the only one that exists within each human organism. Other information processing systems are assumed to be non-conscious rather than merely inaccessible from the perspective of this consciousness. We say information goes from being unconscious to conscious, rather than entertain the possibility that the information may instead be moving between mutually inaccessible spheres of consciousness. The living mirror theory, however, holds that multiple systems within a multicellular organism are associated with consciousness.

The conscious system one identifies with is the one that contains the self model and has access to the neural machinery for language (Metzinger, 2009). It is therefore the system that human researchers have used when approaching the problem of consciousness. By asserting itself to be the only conscious system in the human organism, this particular system has misdirected the search to understand the physical basis of consciousness. It has become instead the search to understand the functioning of this one particular system, which is thought to be distributed throughout the human neocortex (Koch, 2004; Koch *et al.*, 2016; Baars, 2005).

The human neocortex is capable of great information processing feats and, as a result, can communicate about its conscious states. Since this conscious system is so complex, its complexity is widely assumed to be related to its being conscious (Tononi and Edelman, 1998). This assumption is unjustified, however, as complex information processing is a requirement for any conscious system to come under our consideration in the first place, not because it is required for

consciousness, but because it is required for communicating about consciousness. Similarly only systems with language can communicate to us that they are conscious, which has led many to believe that language is somehow essential for conscious experience (Fuster, 2015), rather than essential for reporting on conscious experience.

Is it reasonable to entertain the possibility that other conscious systems could exist within you as an organism that the system you identify with does not have access to? The Nobel prize-winning split-brain experiments of Roger Sperry (1961; 1968) demonstrated several decades ago that it is possible for multiple conscious systems to exist within the human organism. When the corpus callosum, the fibre tract that connects the two hemispheres of the brain, is cut during surgery, two subjectivities can be found to exist in the patient (Sperry, 1968). Information in the left visual field is relayed to the right hemisphere which can report what it has seen using the left arm which it controls. The opposite is true for the left hemisphere which often has the added communicative skill of language production, due to the common left lateralization of the relevant neural circuits. Given the private nature of consciousness we can never be certain that each hemisphere is indeed conscious. If one is not a solipsist, however, and grants consciousness to other humans based on the similarity of their brains and behaviours to one's own, this reasoning should extend to each hemisphere of the split-brain patient.

Even without the evidence from split-brain patients, the possibility of multiple islands of consciousness within an organism is as consistent with the observable data as the existence of a single consciousness. This should lead us to question whether a non-linguistic 'unconscious' system like the superior colliculus that guides most of our visual behaviour is truly unconscious or whether it is merely inaccessible from the perspective of the system we call 'I' (Weiskrantz, 1986). The existence of multiple, mutually inaccessible conscious systems in a single organism would be enough to account for why not everything that is represented by the brain seems to enter consciousness — it only appears this way from the perspective of the system that contains the self model.

Could consciousness also exist in systems outside of the brain? The answer to this question depends on whether it is reasonable to believe that consciousness is linked to computation or to special properties of neural material (Penrose, 1989). It is widely accepted that consciousness is intimately related to computation and information processing (Tononi, 2004; Campbell, 2005; Chalmers, 1996; Davenport, 2000).

Information processing and computation are substrate independent (Turing, 1937). From this perspective, there can be nothing special about neural material in and of itself that makes brains capable of 'producing' consciousness; only its computational capacities can be relevant to the question of consciousness. Information and computation are instantiated in all living systems (Farnsworth, Nelson and Gershenson, 2013), and so any link between computation and consciousness necessarily entails the possibility that all living systems might be conscious. Conflating consciousness with cognition leads many to believe that only extremely complex computational machines like the human neocortex are capable of being conscious. Such machines are indeed required for complex cognition, but there is no reason to believe that the kind of complex information processing that is required for cognition should account for the feature of it being like something to be a given system. In keeping with this, consciousness has been found to be unnecessary for most cognitive processes (Bargh and Chartrand, 1999).

The living mirror theory accounts for the presence of consciousness as a feature of living systems that persists for the duration of the system's life. How can the concept of levels of consciousness be understood in this framework? Consciousness, understood as the presence of subjective experience in a system, 'is not gradable it cannot come in degrees' (Bayne, Hohwy and Owen, 2016). When one sleeps, systems that encode one's self model (Metzinger, 2009; Thompson, 2014), and thereby give rise to the experience of being 'you', reduce their activity (Sämann *et al.*, 2011; Hobson and Pace-Schott, 2002; Boly *et al.*, 2008). The experience of being a self may disappear but there is no reason to assume that consciousness itself disappears during sleep states. In fact a range of evidence attests to the fact that it does not (Windt, Nielsen and Thompson, 2016). We can account for the experience of feeling that consciousness disappears during sleep and in anaesthetized states by appreciating that the cognitive models that mediate our experience and memory of being a self stop functioning, giving the impression that there was nothing it was like to be the sleeping or anaesthetized system. Crucially though, this experience of no experience is only had retrospectively, once these models are back online. This *post hoc* experience tells us nothing about whether there was actually anything it was like at the time to be the sleeping system. Understood in this way, we need not entertain the possibility that brain states or chemicals that induce

anaesthesia can alter consciousness itself; they merely change the appearances within it.

7. Explaining 'My' Consciousness

How does the consciousness we typically identify with, the cortical system that has been most investigated, fit in with this picture? The ability of a multicellular system such as the brain to function as a system in its own right can be attributed to the self-organizing dynamics of living systems (Kirchhoff *et al.*, 2018). The brain and the particular consciousness that comes under most study in neuroscience therefore falls under the same framework described here for simpler systems (Friston, 2013). From this perspective, the brain can be seen as a highly specialized organ that performs Bayesian inference in the service of entropy resistance (Clark, 2013; Kersten, Mamassian and Yuille, 2004).

All behaviours exhibited by complex systems such as ourselves are ultimately the result of the physical and informational dynamics of entropy resistance playing out through our organism. The construction of beliefs regarding external reality in order to exist in a world governed by the second law of thermodynamics can be seen as the fundamental function of the brain. All behaviours that we may be conscious of, from eating an apple to fantasizing about the purchase of a large house in the distant future, fit under this scheme. Entropy-resisting systems must take in energy in order to maintain their form, accounting for the observable apple-eating behavioural dynamics of physical systems such as ourselves. The acquisition of resources, shelter, and social status, as well as the ability to model future scenarios, also increase the chance of an organism staying alive and maintaining its form, accounting for the covert house-buying fantasy dynamics described here. These overt and covert behaviours can therefore be seen to be merely more complex forms of the entropy-resisting behaviours that exist in single cells. Even conscious behaviours that may ultimately result in one's destruction fit into this scheme, as there is no guarantee that the system will always hit upon an appropriate solution for entropy resistance.

How do all of these processes come together into this one consciousness in particular? The answer offered here is that they don't. Feeling that they do is a perspectival illusion created by this system containing the self model. It is akin to asking how it is that reality conspires to make your consciousness feel so much more real than

everyone else's. The fact of the matter is that it doesn't. If your cerebellum were given the ability to self-model and communicate, it too would demand that any explanation of consciousness must ultimately account for its specialness.

If the contents of consciousness always relate to events that are external to the conscious system, how is it that you can have conscious experiences of events that occur within yourself? The 'you' in the previous sentence is the conscious system that you identify with, the conscious 'I' system in your brain that contains your self model. The 'yourself' refers to the organism as a whole. Bodily sensations may be internal to the physical body but they are external to the boundary of the conscious 'I' system. Conscious systems can also exist alongside, as well as within, each other. This accounts for how we can become conscious of thoughts. Systems that allow for 'unconscious' intuitive reasoning, for example, can be conceptualized as independently conscious systems, but once the system externalizes the results of its processing through action potential firing, the conscious 'I' system can construct beliefs about the meaning behind the pattern of neural firing. In this way, becoming conscious of thoughts or other cognitive events can be understood as an act of perceiving neural events that are actually external to the conscious 'I' system. Becoming conscious of a memory, for example, can be conceptualized as the cortical consciousness system that we identify with perceiving the output of the hippocampal memory system.

A broad range of experiences exist in this unified sphere of consciousness associated with the self model. There is not only the experience of smell, for example, but also the appearance of the self and of linguistic thought that can introspect and interrogate sensory experience. Is it the case that a low-level sensory experience is being combined here with a high-level cognitive experience? While sensory perception evidently requires less complex neural machinery than linguistic thought, as evidenced by the scarcity of language in the animal kingdom but the abundance of olfaction, this has no bearing on the extent to which they confer survival value, as is also evidenced by this comparison. As a result, their seeming levels of complexity do not relate to any hierarchy within the space of conscious perception. For the organism as a whole, the consciousness that contains the self model must also contain beliefs about the qualities of the external world, as well as beliefs about the cognitive capabilities of the organism itself that it can use in the service of navigating its environment in order to survive. Beliefs about external objects and one's

cognitive capacities are highly relevant to this task whereas consciously perceiving the state of one's appendix is not often relevant, accounting for why not all possible sources of information are included in the consciously perceived scene. Consider the example of smelling food of questionable age. There is the perception of the smell and the appearance of linguistic thought interrogating the percept. The belief of the system is that there is an organism with the capacity for symbolic thought and a sense of smell. The symbolic thought is triggered by the total context of the organism and enters consciousness along with the olfactory information. Frontal areas whose activity correlates with this particular sphere of consciousness (Koch, 2004; Koch *et al.*, 2016; Baars, 2005) may receive action potentials from the linguistic broca's area and olfactory cortex simultaneously and use both in constructing beliefs about the scene that will be relevant for the organism's survival.

8. Consciousness is Not Fully Substrate Independent

Computation and information processing do not depend on substrate; one can compute with machines as well as with biological matter (Turing, 1937). If consciousness is held to be synonymous with particular abstract computations, then it should be possible to create simulated consciousnesses on computer architectures in which information is encoded in the binary states of transistors (Bostrom, 2003). Consciousness is described here as a computational feature of bounded, entropy-resisting systems at the level of statistical physics, however, not as an abstract computation that can float free of the physical dynamics that instantiate it.

If one were to run a simulation of the appropriate dynamics or informational properties, would that system become conscious? The answer is a definite 'no' as the internal and external aspects of the simulation would only exist at the conceptual level in human minds, not at the level of statistical physics. Without a human mind present, there are merely transistors switching on and off. If one were to program a self-driving car using the same principles that the living mirror theory is based on would the car be conscious? The answer is again 'no'. The physical transistors that encode the relevant information do not actually have an interior vs. exterior aspect, from the perspective of statistical physics. They are part of the non-conscious physical reality that is governed by the tendency towards increased entropy. There is no system there for us to begin to entertain the

possibility of its being conscious. If the car were constructed with a cellular structure so that its dynamics were sufficiently embedded in the physical world so that the computational features associated with entropy resistance could emerge, then it would be predicted to be conscious. The crucial difference between the typical self-driving car and the living system is its boundedness at the level of statistical physics. If consciousness is understood to be an informational feature that emerges at the level of statistical physics in relation to bounded, entropy-resisting systems, electrical circuits can be seen to exhibit no such dynamics. The metal of the transistor is continuous with the air that oxidizes the metal, there is no system there that can be defined statistically in the first place so the question of whether such a system is conscious is meaningless. The difference between the transistor and the neuron is that the neuron is a system at the level of statistical physics and can be part of a larger, self-organizing system that exhibits the same dynamics whereas the transistor is not. Consciousness therefore cannot be instantiated in simulations (Bostrom, 2003), artificially intelligent systems (McDermott, 2007), or complex information processing systems such as the internet (Koch, 2014; Tononi and Koch, 2015; Tononi, 2011) that are based on the information processing architectures that currently exist. The living mirror theory offers a framework, however, for the development of conscious artificial intelligence through the design of systems that sufficiently imitate the dynamics of living systems.

9. Conclusion

The living mirror theory holds that consciousness is a necessary and intrinsic feature of all living systems. In a universe moving towards ever increasing entropy, living systems manage to exist and survive by maintaining a boundary with their environment. This boundary maintenance is computationally equivalent to Bayesian inference, with the internal states of the living system instantiating properties of the external world that account for the sense-data it receives. The living mirror theory holds that through this computational interaction between a living system and its environment, implicit statistical features of the physical, quantitative world get made explicit in informational terms, bringing into existence beliefs in qualities. Surviving through boundary maintenance can be seen from this perspective to be synonymous with the emergence of consciousness, as without consciousness, and the associated physical entropy-resisting

dynamics that give rise to it, living systems would be unable to respond to their environment and thereby exist over time. With the advent of life, the complex structure of the material world became reflected in informational terms through the dynamics of living systems, giving rise to a framework of beliefs in qualities. This framework of beliefs is consciousness, where all experience appears, like reflections in a mirror.

Acknowledgments

This work would not have been possible without the support of Daniel Bendor at the Institute for Behavioural Neuroscience (IBN). I would like to thank my colleagues at the institute and everyone else who provided feedback, especially Robin Mazumder, Soraya Dunn, Fabian Peters, Vanessa Carr, James Street, and Will de Cothi. I am especially grateful to Rebecca Stellato for her extensive feedback and support during the formulation of these ideas and for consistently rolling her eyes whenever anyone doubted the possibility of plant consciousness.

References

Baars, B.J. (2005) Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience, *Progress in Brain Research*, **150**, pp. 45–53.

Bargh, J.A. & Chartrand, T.L. (1999) The unbearable automaticity of being, *American Psychologist*, **54** (7), pp. 462–479.

Bayne, T., Hohwy, J. & Owen, A.M. (2016) Are there levels of consciousness?, *Trends in Cognitive Sciences*, **20** (6), pp. 405–413.

Boly, M., Phillips, C., Tshibanda, L., Vanhaudenhuyse, A., Schabus, M., Thanh Dang-Vu, T., Moonen, G., Hustinx, R., Maquet, P. & Laureys, S. (2008) Intrinsic brain activity in altered states of consciousness: How conscious is the default mode of brain function?, *Annals of the New York Academy of Sciences*, **1129** (1), pp. 119–129.

Bostrom, N. (2003) Are we living in a computer simulation?, *The Philosophical Quarterly*, **53** (211), pp. 243–255.

Campbell, J. (2005) Information processing, phenomenal consciousness, and Molyneux's question, in Bermúdez, J.L. (ed.) *Thought, Reference, and Experience: Themes from the Philosophy of Gareth Evans*, pp. 195–320, Oxford: Oxford University Press.

Chalmers, D.J. (1995) Facing up to the problem of consciousness, *Journal of Consciousness Studies*, **2** (3), pp. 200–219.

Chalmers, D.J. (1996) *The Conscious Mind: In Search of a Fundamental Theory*, New York: Oxford University Press.

Chalmers, D.J., *et al*. (2003) Consciousness and its place in nature, in Stich, S. & Warfield, T.A. (eds.) *Blackwell Guide to the Philosophy of Mind*, pp. 102–142, Oxford: Blackwell.

Clark, A. (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science, *Behavioral and Brain Sciences*, **36** (3), pp. 181–204.

Damasio, A.R. (1989) Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition, *Cognition*, **33** (1–2), pp. 25–62.
Davenport, D. (2000) Computationalism: The very idea, *Conceptus-Studien*, **14**, pp. 121–137.
Farnsworth, K.D., Nelson, J. & Gershenson, C. (2013) Living is information processing: From molecules to global systems, *Acta Biotheoretica*, **61** (2), pp. 203–222.
Friston, K. (2013) Life as we know it, *Journal of the Royal Society Interface*, **10** (86), art. 20130475.
Friston, K. (2018) Am I self-conscious? (Or does self-organization entail self-consciousness?), *Frontiers in Psychology*, **9**, art. 579.
Friston, K., Kilner, J. & Harrison, L. (2006) A free energy principle for the brain, *Journal of Physiology — Paris*, **100** (1–3), pp. 70–87.
Fuster, J. (2015) *The Prefrontal Cortex*, New York: Academic Press.
Hobson, J.A. & Pace-Schott, E.F. (2002) The cognitive neuroscience of sleep: Neuronal systems, consciousness and learning, *Nature Reviews Neuroscience*, **3** (9), art. 679.
Hohwy, J. (2013) *The Predictive Mind*, New York: Oxford University Press.
Kersten, D., Mamassian, P. & Yuille, A. (2004) Object perception as Bayesian inference, *Annual Review of Psychology*, **55**, pp. 271–304.
Kirchhoff, M., Parr, T., Palacios, E., Friston, K. & Kiverstein, J. (2018) The Markov blankets of life: Autonomy, active inference and the free energy principle, *Journal of the Royal Society Interface*, **15** (138), art. 20170792.
Kirk, R. & Squires, R. (1974) Zombies v. materialists, *Proceedings of the Aristotelian Society, Supplementary Volumes*, **48**, pp. 135–163.
Koch, C. (2004) The quest for consciousness, *Engineering and Science*, **67** (2), pp. 28–34.
Koch, C. (2014) Is consciousness universal, *Scientific American Mind*, **25**, pp. 26–29.
Koch, C., Massimini, M., Boly, M. & Tononi, G. (2016) Neural correlates of consciousness: Progress and problems, *Nature Reviews Neuroscience*, **17** (5), art. 307.
Levine, J. (1983) Materialism and qualia: The explanatory gap, *Pacific Philosophical Quarterly*, **64** (4)pp. 354–361.
McDermott, D. (2007) Artificial intelligence and consciousness, in Thompson, E., Moscovitch, M. & Zelazo, P.D. (eds.) *The Cambridge Handbook of Consciousness*, pp. 117–150, Cambridge: Cambridge University Press.
Metzinger, T. (2009) *The Ego Tunnel: The Science of the Mind and the Myth of the Self*, New York: Basic Books.
Nagel, T. (1970) Armstrong on the mind, *The Philosophical Review*, **79** (3), pp. 394–403.
Nagel, T. (1974) What is it like to be a bat?, *The Philosophical Review*, **83** (4), pp. 435–450.
Pearl, J. (2014) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Amsterdam: Elsevier.
Penrose, R. (1989) *The Emperor's New Mind*, Oxford: Oxford University Press.
Revonsuo, A. (1999) Binding and the phenomenal unity of consciousness, *Consciousness and Cognition*, **8** (2), pp. 173–185.

Rosch, E., Varela, F. & Thompson, E. (1991) The embodied mind, *Cognitive Science and Human Experience*, Cambridge, MA: MIT Press.

Sämann, P.G., Wehrle, R., Hoehn, D., Spoormaker, V.I., Peters, H., Tully, C., Holsboer, F. & Czisch, M. (2011) Development of the brain's default mode network from wakefulness to slow wave sleep, *Cerebral Cortex*, **21** (9), pp. 2082–2093.

Schrödinger, E. (1944) *What is Life? The Physical Aspect of the Living Cell and Mind*, Cambridge: Cambridge University Press.

Solms, M. (2018) The hard problem of consciousness and the free energy principle, *Frontiers in Psychology*, **9**, art. 2714.

Solms, M. & Friston, K. (2018) How and why consciousness arises: Some considerations from physics and physiology, *Journal of Consciousness Studies*, **25** (5–6), pp. 202–238.

Sperry, R.W. (1961) Cerebral organization and behavior, *Science*, **133** (3466), pp. 1749–1757.

Sperry, R.W. (1968) Hemisphere deconnection and unity in conscious awareness, *American Psychologist*, **23** (10), pp. 723–733.

Thompson, E. (2014) *Waking, Dreaming, Being: Self and Consciousness in Neuroscience, Meditation, and Philosophy*, New York: Columbia University Press.

Tononi, G. (2004) An information integration theory of consciousness, *BMC Neuroscience*, **5** (1), pp. 1–22.

Tononi, G. (2011) The integrated information theory of consciousness: An updated account, *Archives Italiennes de Biologie*, **150** (2/3), pp. 56–90.

Tononi, G. & Edelman, G.M. (1998) Consciousness and complexity, *Science*, **282** (5395), pp. 1846–1851.

Tononi, G. & Koch, C. (2015) Consciousness: Here, there and everywhere?, *Philosophical Transactions of the Royal Society B: Biological Sciences*, **370** (1668), art. 20140167.

Turing, A.M. (1937) On computable numbers, with an application to the entscheidungsproblem, *Proceedings of the London Mathematical Society*, **2** (1), pp. 230–265.

Weiskrantz, L. (1986) *Blindsight: A Case Study and Implications*, Oxford: Oxford University Press.

Windt, J.M., Nielsen, T. & Thompson, E. (2016) Does consciousness disappear in dreamless sleep?, *Trends in Cognitive Sciences*, **20** (12), pp. 871–882.

Paper received June 2019; revised February 2020.